Lényegkiemelő módszerek összehasonlítása közlekedési zajban történő beszédfelismerés céljából

Sárosi Gellért[1], Tobler Zoltán[2], Mihajlik Péter[1,2], Fegyó Tibor[1,3]

[1] Távközlési és Médiainformatikai Tanszék,
Budapesti Műszaki és Gazdaságtudományi Egyetem
{sarosi, tobler, mihajlik, fegyo}@tmit.bme.hu
[2] THINKTech Kutatási Központ Nonprofit Kft.
[3] Aitia International Inc.

Kivonat: A gépi beszédfelismerés egyik döntő fontosságú eleme a beszéd akusztikai lényegének kiemelése, különösen a zajos környezetben történő alkalmazásoknál, amely jelen esetben közlekedési zajjal terhelt akusztikai környezetet jelentett. Emiatt helyeztük vizsgálatunk középpontjába a zajtűrő és hagyományos beszédfelismerési lényegkiemelési eljárásokat. A tanítást és tesztelést hat nyelven végeztük el: angol, francia, magyar, német, olasz, spanyol. Teszteléshez a telefonos hálózaton keresztül az utcáról vagy járműből rögzített adatbázist használtunk. Alaprendszerként teszteltük a HTK és a SPHINX eszközkészletben, vagy általunk is implementált Mel Frequency Cepstral Coefficients (MFCC) és Perceptual Linear Prediction (PLP) módszereket. Az újabb módszerek között a Power-Normalized Cepstral Coefficients (PNCC) és a Perceptual Minimum Variance Distortionless Response (PMVDR) szerepel.

1 Bevezetés

Feladatunk közlekedési zajban üzemelő folyamatos beszédfelismerő rendszer összeállítása. A rendszernek hat nyelven kell működnie: angol, francia, magyar, német, olasz és spanyol. A cél: felismerni a nyilvános mobiltelefon-hálózaton érkező hívásokban, hogy a hívók milyen célobjektumot (POI – Point of Interest) szeretnének megtalálni, mint például egy múzeumot, éttermet vagy konkrét címet. A rendszernek legalább a POI-k többségét megbízhatóan fel kell ismernie annak ellenére, hogy az utcán sétálva, vagy valamilyen járműben utazva a beszédkörnyezet legtöbbször zajjal terhelt.

2 A lényegkiemelők

Az automatikus beszédfelismerés kritikus lépése a lényegkiemelés, hiszen ekkor alakítjuk át a beszédet a gép számára feldolgozható lényegvektorok sorozatává. Emiatt helyeztük kísérleteink középpontjába különféle lényegkiemelő eljárások vizsgálatát.

A *Mel Frequency Cepstral Coefficients* (MFCC) egy elterjedten alkalmazott módszer, sokféle implementációja létezik, ezek közül hárommal foglalkoztunk. Az egyik a HTK (Hidden Markov-Model Toolkit) [10] nevű, rejtett Markov-modellek építésére és manipulációjára alkalmas eszközkészlet. A munkánk során részint a beépített lényegkiemelő eszközöket, részint az akusztikus modelltanító és -kiértékelő eszközöket használtuk fel. A másik a SPHINX [1] nevezetű, kifejezetten beszédfelismerésre készült rendszer, ennek csupán a lényegkiemelő részét használtuk fel. A harmadik a saját implementációnk, mely a Voxerver[1] nevezetű felismerő szoftver része. Mindhárom módszer magja a Mel-szűrőbank és a logaritmikusamplitúdó-kompresszió.

Zaj szempontjából robosztusabb megoldást kínálhat a *Perceptual Linear Prediction* (PLP) módszer [2], mely lineáris predikciót (LP) használ a beszéd spektrális burkolójának előállításához. A perceptualitást a Bark-szűrés és – a hallás frekvenciával változó érzékenységét követő – azonos hangosságú előkiemelés adja.

Az újabb módszerek között szerepel a *Perceptual Minimum Variance Distortionless Response* (PMVDR) [9], mely szűrés helyett egy paraméterezhető ún. frekvenciahajlítást (frequency bending) alkalmaz, a LP-együtthatókból pedig MVDR-spektrumot, egy felső spektrális burkolót számít.

Szintén új módszer a *Power-Normalized Cepstral Coefficients* (PNCC) [3], amely ún. Gammatone-szűréssel [7], teljesítményeltolással és exponenciális amplitúdóösszenyomással reprezentál egy zajrobosztus lényegkiemelést.

3 A kísérleti környezet

Tanítási célokra a SpeechDat [8] adatbázist használtuk, mely az általunk vizsgált nyelveken, egyenként 500-5000 beszélőtől származó, a vezetékes és a mobilhálózaton keresztül is rögzített felvételeket tartalmaz. Ez alól kivétel a magyar nyelv, amelynél az akusztikai modellek tanításához az MTBA-t (Magyar Telefonos Beszéd Adatbázis) [4] használtuk. Ez 500 beszélőtől tartalmaz felolvasott szöveget szintén a telefonhálózaton keresztül rögzítve, tehát teljességgel SpeechDat-szerű, és felhasználható a kísérleteinkben. Mind a tanító-, mind a tesztadatokra közösen jellemző a 8kHz mintasűrűség, az egy csatorna és a 8 bit A-law kódolás. Az angol nyelv esetén fellépett adatelégtelenségi problémák miatt ezt a nyelvet kivettük a zajtűrési vizsgálatainkból. A teljes adatbázisból 10 órányi adatot használtunk fel a tanításra, mert kísérletünk célja az egymáshoz viszonyított javulások vagy romlások feltérképezése, nem pedig az abszolút legjobb felismerés, ez esetben pedig az adatbázis mérete nem kritikus.

A felismerési teszteket két szakaszban hajtottuk végre. Elsőre verifikációs teszteket végeztünk magyar nyelven, hogy beállítsuk az optimális tanító és tesztelő környezetet. E célból felhasználtunk mintegy 15 percnyi hanganyagot egy magyar nyelvű műsorszóró adó híradójából, és ugyanennyi telefonos hanganyagot. A tesztek lényegi részét a többnyelvű felismerések adták, melyekhez a telefonos hálózaton keresztül az utcáról vagy járműből rögzített, tájékozódási célú kérdésekből és kijelentésekből álló, alacsonyabb jel-zaj viszonyú adatbázist használtunk. Tartalmuk egy-egy POI megta-

[1] Aitia International Inc.

lálásához kapcsolódó kérdés vagy jellemzés, de vannak POI-t nem tartalmazó bemondások is a szófelismerés pontosabb méréséhez.

Miután a tanító adatokon lefuttattuk a lényegkiemelést, fonémaszintű címkézés és flat-start módszer [10] alkalmazásával is Maximum Likelihood módszerrel tanítottunk GMM (Gauss Mixture Model) alapú, szóhatárokon átívelő trifón akusztikai modelleket. A felismerési teszteket a már korábban említett Voxerverrel végeztük, mely egy WFST (Weighted Finite State Transducer) [5] alapú dekóder szoftver. Az akusztikai modell mellett WFST alapú nyelvtanokat használtunk, melyeket a lehetséges kérdező, kérő mondatszerkezetekből és POI-kifejezésekből generáltunk a [6] módszerei szerint.

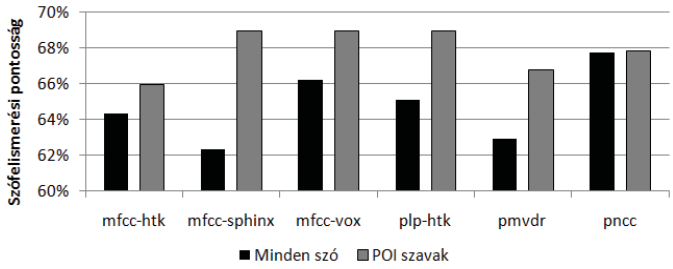
4 Eredmények

Minden kísérletben 39 dimenziós, az energia jellemzőt is magában foglaló lényegvektorokat állítottunk elő. Az akusztikai modelleket alkotó Gauss-függvények számát 10-ben maximalizáltuk. A nyelvenként és módszerenként lefutott legjobb felismerési eredményeket a 1. táblázat tartalmazza, ahol WAcc a szófelismerési pontosságot jelenti, az *All* és *POI* pedig hogy az eredmény minden szóra vagy csak a POI-kra vonatkozik. Az első három sorban a három MFCC változat, a másik háromban a zajrobosztus frontendek pontosságai szerepelnek, kiemelve a nyelvenkénti legjobbat.

1. táblázat: Felismerési eredmények.

WAcc	Francia		Magyar		Német		Olasz		Spanyol	
(%)	All	POI	All	POI	All	POI	All	POI	All	POI
mfcc-htk	56.8	70.3	70.1	57.4	62.6	56.5	62.7	**81.2**	75.4	77.6
mfcc-sphinx	**65.9**	74.4	68.3	60.6	59.5	69.6	57.7	73.9	71.6	65.7
mfcc-vox	**65.9**	70.3	67.5	56.9	**69.3**	63.0	62.1	76.1	77.1	76.1
plp-htk	61.4	**75.2**	71.1	**62.3**	66.3	69.6	**63.7**	70.3	79.8	**85.1**
pmvdr	65.5	74.4	69.7	59.7	62.0	**71.7**	59.3	75.4	68.3	56.7
pncc	64.8	70.3	**71.3**	61.7	67.5	67.4	59.5	68.1	**83.6**	80.6

A legkiemelkedőbb a magyar 71.3%-os szófelismerési arány, melyet viszonylag nagyobb tesztadatbázis mellett sikerült elérni. A spanyol 83.6% is figyelemre méltó, de a kevesebb tesztadat miatt kevésbé megbízható. Az olasz, francia és német adatokon is magas POI-pontosságot értünk el, de általában az összes szó felismerése ettől nem marad el jelentősen.



1. ábra. Felismerési eredmények módszerenkénti átlagai.

A kapott értékekből módszerenkénti átlagot képeztünk, hogy megkeressük a globálisan optimális lényegkiemelő módszert. Ez látható az 1. ábrán. Az átlagosan legjobban teljesítő módszer a PNCC, legmagasabb átlagos szófelismerési pontossággal. Még a PNCC-nél is jobb POI-felismerést adott a HTK PLP és saját MFCC-implementációnk. Az összes szót tekintve már rosszabban teljesítettek, viszont a saját MFCC-implementációnk jobb teljesítményt mutatott a másik két változatnál. A SPHINX MFCC módszere szintén kiemelkedő POI-pontosságot ért el, de az összes szót tekintve a leggyengébben szerepelt. A PMVDR és HTK MFCC átlagosan gyengébben teljesített, bár az eredmények közti eltérések nem jelentősek.

5 Összefoglalás

Öt nyelven készült el egy olyan beszédfelismerő rendszer, amellyel nyelvenként hatféle lényegkiemelési módszer szerint végeztünk kísérleteket. Eredményeink alapján a PNCC teljesített a legjobban, mert sok nyelv esetén a legjobb, vagy ahhoz közeli felismerést adott. Szintén kiemelkedő a Voxerver MFCC és HTK PLP teljesítménye, de átlagban kissé elmaradnak a PNCC-től. Ráadásul a Voxerver MFCC jobban teljesít a másik két implementációnál. Az angol rendszer gondjai kevés fejlesztéssel megoldhatóak, és az abszolút felismerési eredmények is tovább javíthatóak, ha a teljes adatbázist felhasználjuk a tanításhoz, ezt tekintjük a lehetséges folytatás fő irányának.

Köszönetnyilvánítás

Kutatásainkat részben támogatták: OM-00102/2007, OMFB-00736/2005, TAMOP-4.2.2-08/1/KMR-2008-0007.

Bibliográfia

1. CMU Speech Recognition Engine (SphinxTrain 1.0): http://www.speech.cs.cmu.edu/

2. Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. Journal of the Acoustical Society of America Vol. 87 No. 4 (1990) 1738–1752
3. Kim, C., Stern, R. M.: Feature Extraction for Robust Speech Recognition using a Power-Law Nonlinearity and Power-Bias Subtraction. In: INTERSPEECH (2009) 28–31
4. Magyar Telefonos Beszéd Adatbázis: http://alpha.tmit.bme.hu/speech/hdbMTBA.php
5. Mohri, M., Pereira, F., Riley, M.: Weighted Finite-State Transducers in speech Recognition. Computer Speech and Language Vol. 16 No. 1 (2002) 69–88
6. Mozsolics T., Tarján B., Mihajlik P., Fegyó T.: Környezetfüggetlen és sztochasztikus nyelvtanok összehasonlítása többnyelvű gépi beszédfelismerési feladatban. In: VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 203–215
7. Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., Allerhand, M. H.: Complex sounds and auditory images; Auditory and Perception (1992) 429–446
8. SpeechDat(II) telephone network database. http://www.speechdat.org/SpeechDat.html
9. Yapanel U. H., Hansen, J. H.L.: A New Perspective on Feature Extraction for Robust In-Vehicle Speech Recognition In: EUROSPEECH (2003) 1281–1284
10. Young, S., Ollason, D., Valtchev, V. and Woodland, P.: The HTK book. (for HTK version 3.4), March 2009. http://htk.eng.cam.ac.uk